RACV2016

Research and Application in Computer Vision

Shanghaitech China

Hao Ding

Sep. 18-20,2016

摘要

The conference is held by China Computer Federation(CCF) and lasts for three days. I'm glad to took part in the conference and I'll finish my report as soon as possible as I didn't forget the details.

目录

Part I

Posters

Section 1

Sep. 18

Two o'clock in the afternoon on Sept. 18th, we come to the ShanghaiTech University to log in the conference. Then we browsed the posters from different university before dinner. Here I'll introduct several of them that I am intrested in.

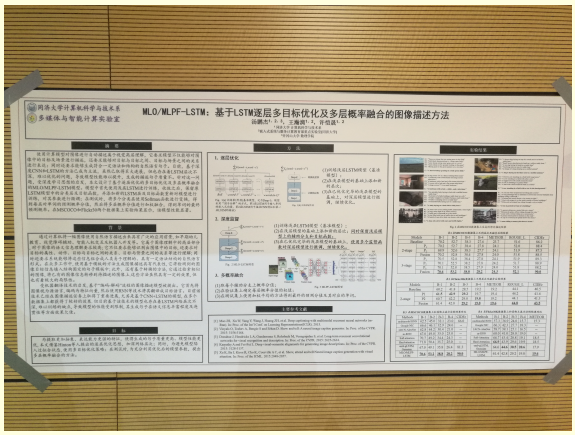

图 1: MLO/MLPF-LSTM: 基于 LSTM 逐层多目标优化及多层概率融合的图像描述方法

Explaination of nouns:

- CNN Convolutional Neural Network.To get the next tier of image by convolutional kernals.
- RNN Recurrent neural network/Recursive neural network. RNN uses all-linked network to get the next image.
- LSTM Long Short Time Memoring. LSTM is a method belongs to RNN. It imitates the human brain to memorize the previous images.

The figure 1 used deep learning to recognize objectives and to semantical analysis image. I firstly understood the main idea in deep learning after reading carfully to the details with the explaination from biaobiao.

The formally way to semantic analysis images is CNN + LSTM model. Though it gets good grades in concluding contents, it's not optimised enouph to get our goal because of its swallow network layer.

The original point in this poster is that the author used hierarchical optimization method to deeper the network. On the other side, he used multiple probability fusion way to avoid overfitting. Multiple probability fusion means to add three arguments named p1, p2, p3 in the network to make the deep neural network can be affected by the last and the swallow layers directly.

We can see that the method works from the data result on the poster.

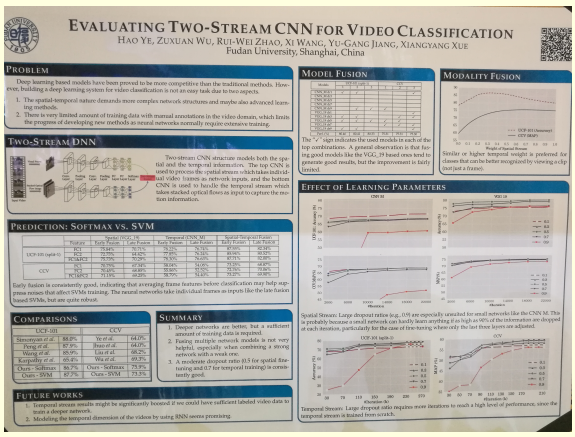

图 2: EVALUEATION TWO-STREAM CNN FOR VIDEO CLASSIFICATION

The figure 2 is a poster aims to add label to videos. It uses two CNN to create the entire neural network. One is to proceed the spacial stream and the other is to proceed the current stream. The spatial stream takes individual video frames as network inputs and the temporal CNN takes stacked optical flows as input to capture the motion information.

In this research, the author innovaionally used ratial connection to build the neural network. Which means to disconnect the joints between the layers by a proportion to avoid overfitting.

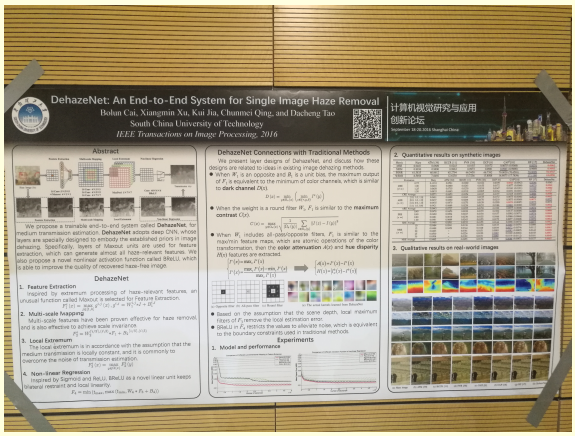

图 3: DehazeNet:An End-to-End System for Single Image Haze Removal

The figure 3 is about dehaze algorithm. I don't understand all of the formulaes, but the method looks great in comparation to the other methods.

The poster 4 is the one most close to my direction. The main point is to matching a sketch to a real object. The classification isn't simply classify what the sketch is. It also considers the

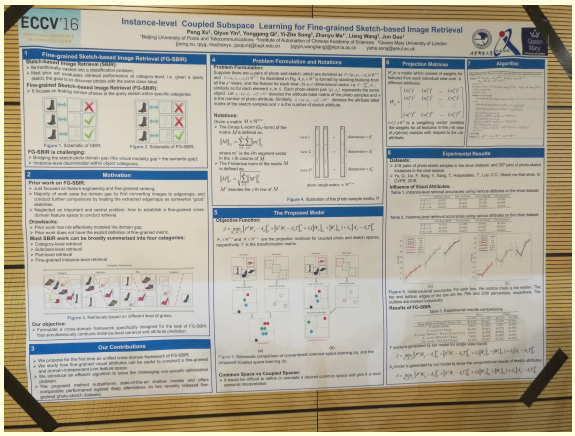

图 4: Instance-level Coupled Subspace Learning for Fine-grained Sketch-based Image Retrieval

outlook features to get its material object.

The author firstly extracts several features from both the sketch and the object. Then he tries to match them in the feature space. The result says the higher k is, the higher accuracy we'll get. The decline string in the picture is because over flitting.

Acctually, the accuracy can't meet the result of deep learning. However, as for a same dataset, the deep learning method has to run for more than one week while this method needs less than one second. Considering the time, the tiny distance in accuracy can be ignored.

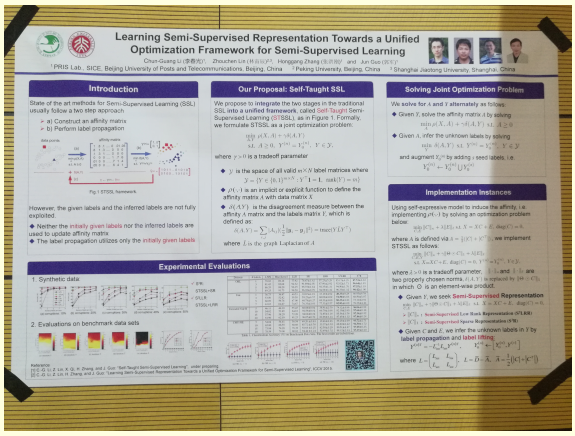

图 5: Learning Semi-Supervised Representation Towards a Unified Optimization Framework for Semi-Supervised Learning

One of the main references of my graduation project is witten by professor ChunGuang Li. Thus I'm grateful to see himself here. The poster mainly uses SRC through semi-supervised learning to recognize human faces in condition of lacking labels.

The kernel of this algorithm is a STSSL framework. The arguments in the framework lets the output of images each time to get more similar conscequence by the scanties of images with label. Then we label the found ones so that we can get more and more labeled images.

The experiment runs on the four classical datasets named ORL, Yale, Extended Yale B and

CMU PIE. The more we run the framework, the more labels we can get, and the higher accuracy we will reach. To my surprise, with such tiny sample size, the result accuracy is so close to the accuracy I got by SRC with more than 50% testing samples.

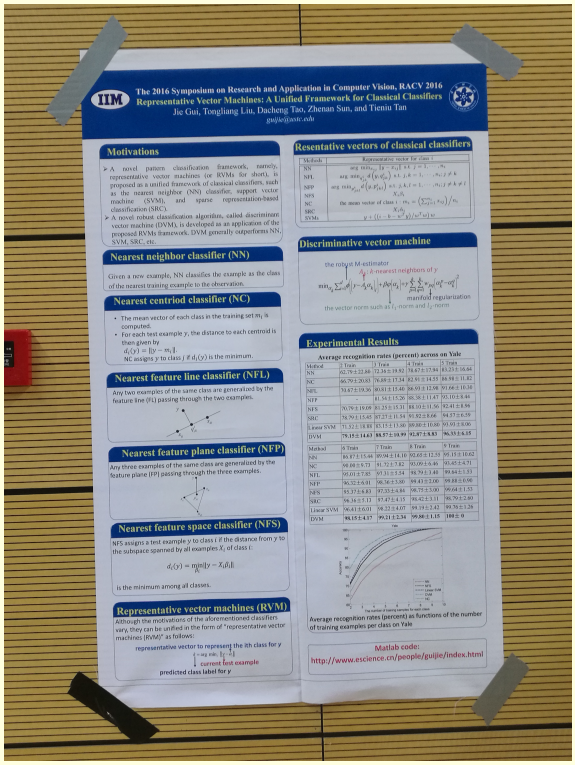

图 6: Representative Vector Machines:A Unified Framework for Classical Classifiers

The figure 6 is the most understandable one for me and the easier one in teacher Yu's opinion. It introduces nearest neighbor classifier(NN), nearest centriod classifier(NC), nearest feature line classifier(NFL) and nearest feature space classifier(NFS). Then the author summarized all of the classifiers to one formular. He created a new operator based on NN and SVM as well.

Figure 7 is written by the same author as the last one. But I didn't follow it carfully. It's about feature selection.

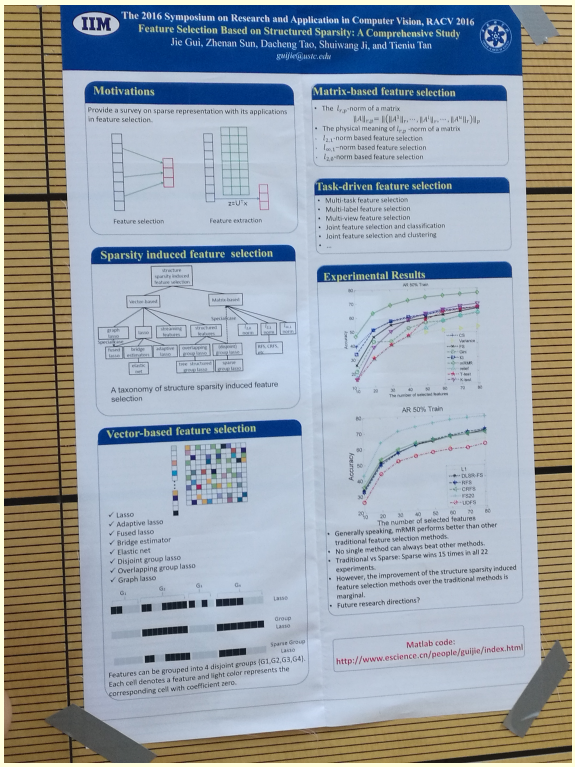

图 7: Feature Selection Based on Structured Sparsity:A Comprehensive Study

Part II

Reports

Section 2

Sep. 19

2.1 Lihi Zelnik-Manor

Lihi Zelnik-Manor is one of the two forein professors. The title of her lecture is Separating the Wheat from the Chaff Visual Data.

Professor Lihi's work can be separated to two main work. Firstly she tried to select the best picture which can best describe the images. The second work is to extract the main pixels from the whole image.

Lihi explained that when we look at a picture or a video, we just focus on one object at once. We always focus on the pattern distinctness, color distinctness, organization priors and objects or

图 8: Separating the Wheat from the Chaff Visual Data

faces.

Thus it's important to correctly extract the most important object from a image. Here are some results before:

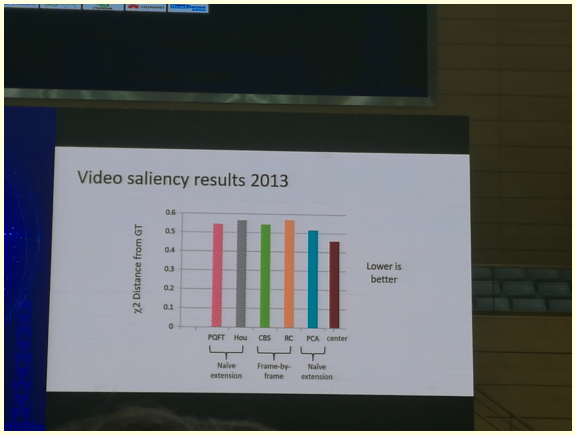

图 9: Separating the Wheat from the Chaff Visual Data

Lower is better and the video saliency methods are no better than just use the center. It's interesting isn't it?

So Lihi then try to conclude the real focus of a video or an image. The results are as follows:

This seems better.She then claimed that video is different from image. Video is not just a set of images and video saliency can be predicted.

2.2 JiaYa Jia

JiaYa Jia mainly introduced his works during these years, which is familiar with me that I've seen the work in our lab for more than one time. The name of his topic is Computer Vision that Mimics and Surpasses Human Ability.

He said that the computer vision can be divided to three main level: Low-level vision, Middle-

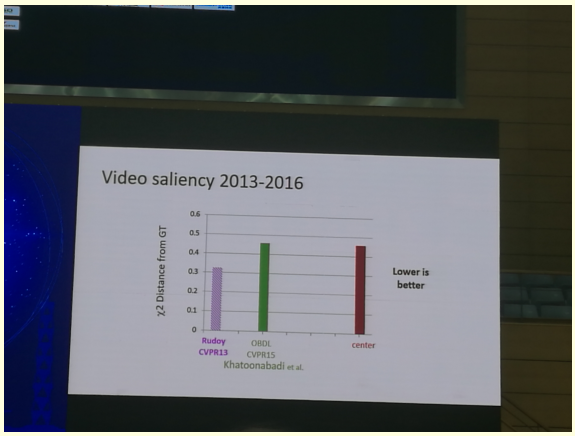

图 10: Separating the Wheat from the Chaff Visual Data

level Vision and High-level Vision. He also did lots of work in photo editing and enhancement, rolling guidance filter for deblurring and vision and language.

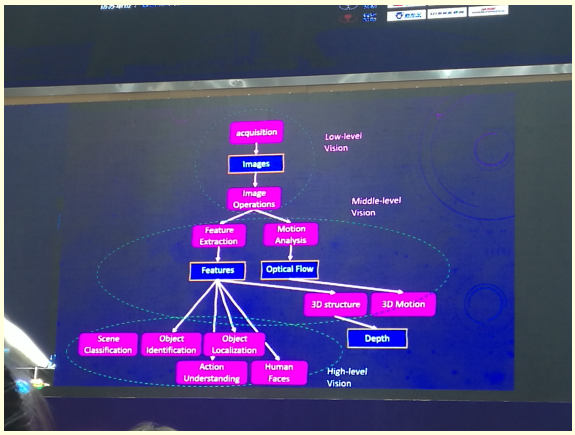

图 11: Computer Vision that Mimics and Surpasses Human Ability

Professor Jia explained that many vision tasks require very detailed pixel processing hat is beyond human ability. Besides, high-level vision needs large data and good learning models-learning human ability is also challenging.

2.3 Specially invitaion

Lin Mei mainly discussed the application of AR and VR.

Augmented Reality(AR) is a live direct or indirect view of a physical, real-world environment whose elements are augmented(or supplimented) by computer-generated sensory input such as sound, video, graphics or GPS data.

Lin Mei refered to Pokemon Go, which is my most expect game. He said the game isn't approached to the kernel of AR but attracts millions of fans. So the AR field surely has bright future.

图 12: AR and Games

Prof. Xi Li is the second speaker of the specially invitation. He showed us an interesting picture contains different fields in computer learning.

图 13: AI-Driven Visual Understanding and Computing

Xi Li wants to let the machine ”think” like a human brain and understand the vision data.

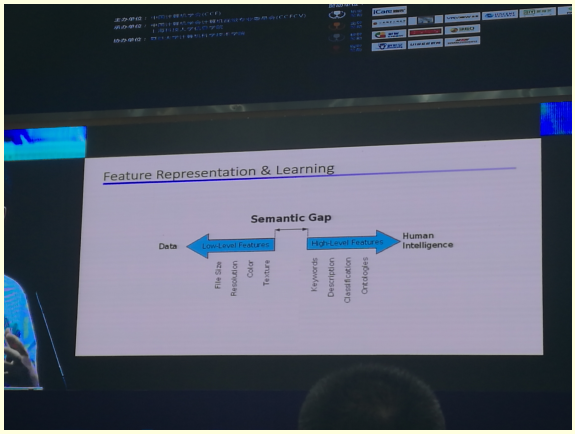

图 14: AI-Driven Visual Understanding and Computing

Then he introduced GoogleNet. I learned that its a popular deep learning network recently. He also mentioned object detection, machine intelligence, object tracking and several other projects.

2.4 ShiGuang Shan

Prof. ShiGuang Shan made his speech of “基于深度学习的人脸检测与识别进展及开放问题”.

He primarily introduced how to recognize human face under complex condition and the progress we have obtained in combining traditional method and deep learning.

Surveillance Human Action Dataset(SHAD) is a video dataset to monitor pedestrian actions. The dataset contains more than 300 videos with complete label.

I think his work is extraordinary useful because it aimes at solving public security problems.

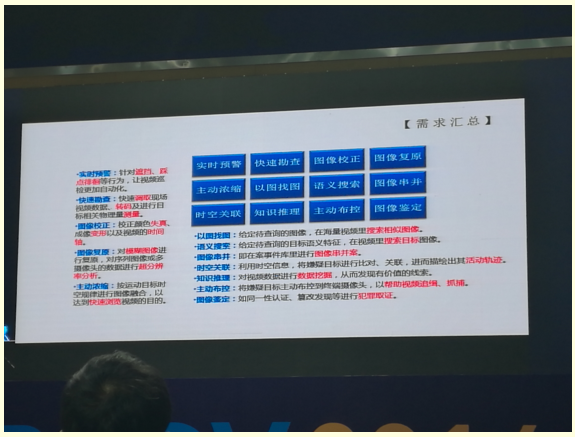

图 15: Public Security & Machine Learning

Section 3

Sep. 20

3.1 Long Quan

Long Quan is a Professor of the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology(HKUST). The title of his lecture is Mapping the World with Drones.

At the beginning, he put up with several questions: What are visual features? Where is the camera? What is the depth? What is segmentation and object recognition?

Computer Vision has three stages:

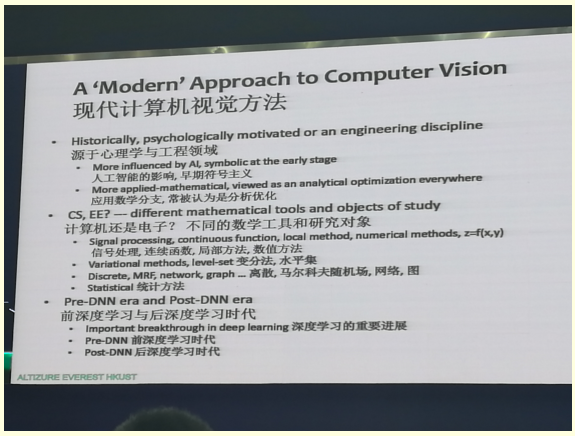

图 16: Three stages of Computer Vision development

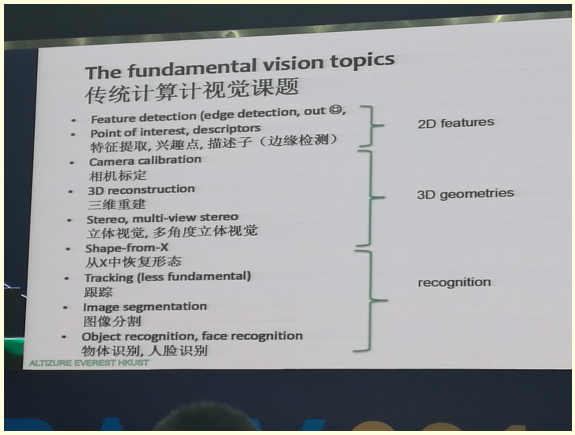

图 17: A 'Modern' Approach to Computer Vision

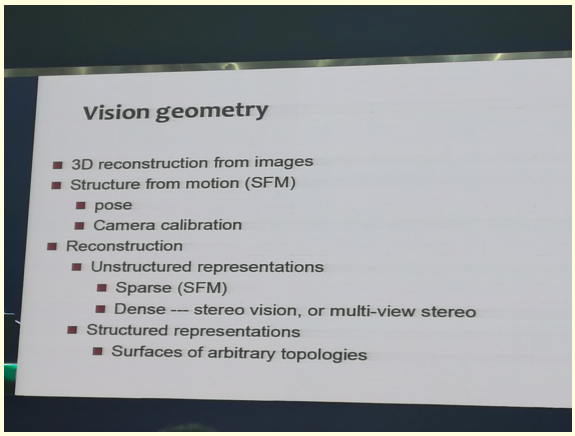

图 18: The fundamental vision topics

He concluded that Computer Vision is fast evolving. His lab is slightly leading at the world stage, and he'll continue to innovate based on their HKUST worldclass lab.

Section 4

conclusion

All of the speechers made great speeches for us.

The first thing is that I deeply feel that I really lack of professional knowledge of my major. And I'm now eager to develop my english ability especially the listening part. My plan is to use english more often. I'll try to read more and watch more english files and videos.

The second thing is that I finally touched the best professors and the best researches in my field. That's excited and will stimulate my study.